学校编码：10384　　分类号_____密级_____
学号：X2013230754　　UDC_____

工　程　硕　士　学　位　论　文

基于大数据分析的 P2P 信贷风险管理系统的设计与实现

Design and Implementation of Peer to Peer Credit Risk Management System Based on Big Data Analysis

郑　旭

指　导　教　师：　林　坤　辉　教　授
专　业　名　称：　软　件　工　程

论文提交日期：　2015　年　4　月
论文答辩日期：　2015　年　5　月
学位授予日期：　　年　　月

指　导　教　师：_____
答辩委员会主席：_____

2015 年 4 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，

于　　年　月　日解密，解密后适用上述授权。

（ √ ）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘 要

作为互联网金融典型代表，P2P 网络借贷引起社会各方面的高度关注。同时受经济全球化和全球信息化、人类社会发展和需求多样性、云计算和物联网技术深化应用等方面的影响，“大数据”(bigdata)已经成为 IT 领域和互联网上反复提及的热词。国内 P2P 网贷行业面临的系统风险、信用风险、平台风险等正在不断积累、扩大，如何有效防范 P2P 网络借贷风险迫在眉睫。本文从国内典型的 P2P 平台拍拍贷中抓取实际有效的数据，经过本系统分析得到结果，希望能够对 P2P 行业具有一定的指导作用。

基于大数据分析的P2P信贷风险管理系统通过开发网络爬虫实时与被采集网站连接，自动使用配置好的已注册帐号通过模拟人工登录的方式登录网站，获取了较为全面的信息，及时的并行抓取网站的数据，并且同步的存入数据库服务器中，提高了采集的效率降低了采集数据的成本和时间，同时保证了系统采集信息的实时有效性，最后根据分析结果推测网络借贷中哪些因素会对借款人发生违约产生影响，并且将分析的结果通过系统显示出来。

本文探讨了并行爬虫程序和基于 Spring.Net 框架开发的 P2P 信贷风险管理系统，基于微软的.NET 平台（C#语言），后台使用 SQL Server 作为数据库服务器。在功能上分为三个主要的模块：爬虫参数配置，账户管理配置，数据抓取分析。其中以数据抓取为着重点。爬虫参数配置包括爬虫的开始时间、结束时间、数量、范围等基础参数设定；账户管理配置主要定义了爬虫操作管理员和数据分析人员的各个操作；数据抓取分析是系统的核心业务部分，涵盖了数据抓取，数据存储，数据分析，数据导出等工作。在应用软件开发的过程中以软件工程为指导思想，严格遵守软件工程学的各个原则，遵循大数据理论思想，采用统计学定性分析与定量研究相结合的方法，总结出其中存在的问题并且通过分析指出影响借款人违约的关键因素，以期对中国目前正火热发展的 P2P 网络小额信贷行业的风险监管起到一定的借鉴意义。

关键词：金融信息化；数据采集；大数据

Abstract

P2P network lending, a typical representative of internet finance, has attracted a lot of public attention. Scandals such as causing investors penniless is not uncommon due to the low entry threshold, legal lag, lack of supervision and incomplete personal credit system of the P2P network lending industry. Domestic P2P network lending industry is now facing accumulating and expanding systematic risk, credit risk and platform risk. The situation, therefore, calls for effective methods to prevent the risks. “Big data” has become a hot word in the IT field and constantly used on the internet thanks to the influences brought by the globalization of economics and information, development of the human society, diversification of demands, and deep application of cloud computing and internet technology. People now use the term to define and describe the massive data generated from various areas of the human society in the information explosion age.

This system through the development of real-time web crawler connection with the web site, automatically configured manually by simulating a registered account login log website for more comprehensive information in a timely manner parallel to crawl data website and stored in the database server synchronization, improved collection efficiency and reduce the cost and time of data collection, while ensuring the effectiveness of the system in real-time information collection, finally figured out which factors would borrow the network according to the analysis results borrower default occurs impact, and the results will be analyzed by the system is displayed.

This dissertation discusses the parallel crawlers and credit risk management system which used Spring.Net framework , based on Microsoft's .NET platform (C #), using SQL Server as the database backend server. Functionally divided into three main modules: reptiles parameter configuration, account management configuration, data analysis crawl. Which is the focus of the data fetch. Reptile parameters including the start time reptiles, end time, number, scope and other basic parameter settings; account management configuration defines the various operations mainly reptiles operating personnel administrators and data analysis; data analysis is the core business grab part of the system, covering data capture, data storage, data analysis, data export and so on. In the process of application development software engineering as the guiding ideology, compliant with the principles of software engineering, theoretical ideas follow big data,

using statistical analysis of qualitative and quantitative research method of combining, which summed up the problems and by analysts pointed out that the key factors affecting the borrower defaults. With the hope of offering guidance to practices of the P2P network microfinance, the dissertation at the end summarises the existing problems and points out the key factors that lead to the irregularities of borrowers.

Key Words: Financial Information; Data Acquisition; Big Data

目 录

CONTENTS

第一章 绪论

1.1 研究背景及意义

P2P 网络借贷的英文全称为“Online Peer-to-Peer Lending”，指不以银行等传统金融机构为中介，借贷双方直接经过互联网上的网贷平台完成交易的无担保借贷，我国银监会和小额借贷联盟将给其定义为人人贷。

P2P 网络借贷的发展： P2P 网络小额信贷基于互联网，并且利用了现有的互联网技术对传统的小额信贷进行了创新式发展，从而有效的解决了传统小额信贷供给机构(如农村信用社、银行等)的资金供给不足、投资者信息不对称、投资起点高、风险高等方面难题，不仅是一种新的民间金融模式，还是一种创新型公益模式。P2P 网络借贷弥补了传统金融机构无法顾及的小微个体的资金需求，对小微企业、个人融资以及社会公益扶贫意义重大。

P2P 网络借贷与传统银行机构的区别：新型的 P2P 网络借贷不同于传统银行机构，不同之处在于投资者与借款人直接发生借贷关系，实现了资金方面的直接融通，P2P 网络借贷平台起到信息展示、供需对接等中介作用。二者的区别如图 1-1 所示。

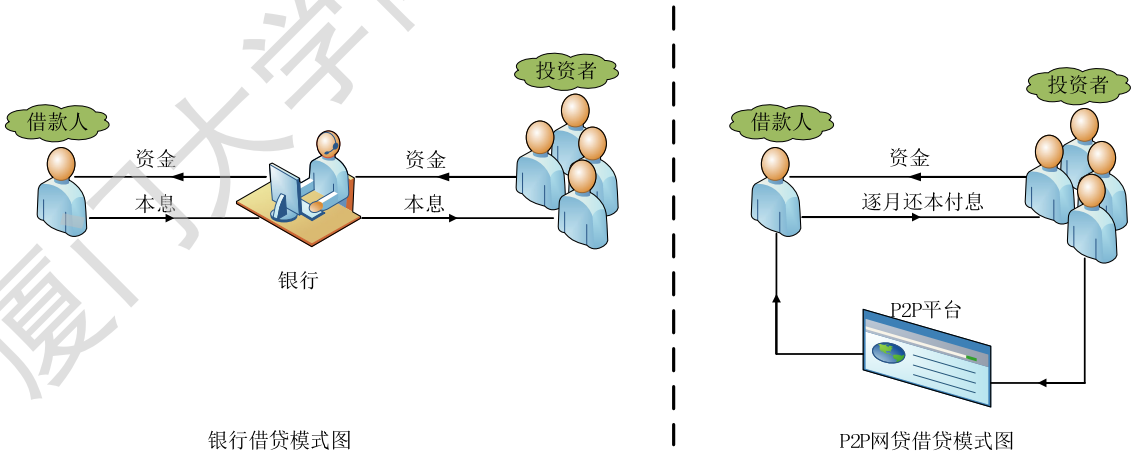

图 1-1 银行借贷与 P2P 网络借贷模式对比

大数据(big data) 大数据是用来描述指数增长和可用性的数据，这样的数据包括两种，一种是结构化的数据，另外一种是非结构化的数据。因为互联网已经形成，

所以大数据将会在经济，社会等领域越来越重要。为什么这样说呢？因为更多的数据可能会使分析更加的准确，而更准确的分析可能会使得人们在决策的时候更加自信，并且获得更高的运营效率，同时降低成本，并且减少风险。2011 年，麦肯锡公司对全世界大数据的分布作了一个研究和统计，中国 2010 年新增的数据量约为 250PB，而欧洲约为 2000PB，美国约为 3500PB，大数据已经深深地充斥了人类经济社会的许多角落[1]。2001 年行业分析师道格·莱尼（现在 Gartner 公司）就阐述了大数据具有 4 大 V 的特点：Volume、Velocity、Variety、Veracity。广义范围上的大数据可分为五种类型：Web 网络社交媒体的数据、机器对机器（M2M）的数据、海量的交易数据、生物计量学的数据和各种人工生成的数据。在各行各业中随处可见因数量、速度、种类和准确性结合带来的大数据问题。

1.2 国内和国外研究现状

国外 P2P 网络借贷始于 2005 年 3 月，发展了约 10 年的时间。P2P 网络借贷起源于 2006 年“诺贝尔和平奖”得主尤努斯教授（孟加拉国）首创的 P2P 小额借贷。随着互联网的快速发展和普及，P2P 小额借贷逐渐与互联网结合起来，并由单一的“线下”模式转变为“线下与线上”并行的模式或者“纯线上”模式，P2P 网络借贷平台应运而生。目前国外学者对 P2P 网络借贷风险的研究，主要集中在信息不对称导致逆向选择和道德风险，社会资本和网络社区对降低违约风险的作用等方面。Sven and Fabian(2009)[4]通过对一 P2P 平台的超过 14000 条的真实借贷记录的实证研究，指出由群组的组长负责监测还款人还款，有助于减少信息不对称问题，提高借款人的信用条件。Freedmanand Jin(2008)[3]以 prosper 为研究对象，发现部分社交网络变量传递的关于借款人的“软信息”在一定程度上补充了网站上部分“硬信息”的遗漏，从而起到降低违约风险的作用。Bergerand Gleisner(2009)[4]研究认为，平台上的众多参与者，如普遍存在的群组关系及每个群组的组长，起到了类似银行等金融中介的作用，对借款人的贷款行为起到了事前的考察、审批以及事后的监督等作用，从而降低了违约比率。Greinerand Wang(2009)[5]认为社会资本确实在 P2P 借贷中起到了一定的作用，但是不同人的受益程度可能有所不同，受益程度最大是风险相对较高的借款人，这部分人群可以通过合理地提高社会资本，从而获得更好的资金支持。